applied sciences

MDPI

Article

A New Partitioned Spatial–Temporal Graph Attention Convolution Network for Human Motion Recognition

Keyou Guo, Pengshuo Wang *, Peipeng Shi, Chengbo He and Caili Wei

School of Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China
* Correspondence: 2130062115@st.btbu.edu.cn

Abstract: At present, human action recognition can be used in all walks of life, because the skeleton can transmit intuitive information without being affected by environmental factors. However, it only focuses on local information. In order to solve these problems, we introduce a neural network model for human body recognition in this paper. We propose a model named NEW-STGCN-CA. The model is based on a spatial–temporal graph convolution network (ST-GCN), which contains a new partition strategy and coordination attention (CA) mechanism. By integrating the CA attention mechanism model, we enable the network to focus on input-related information, ignore unnecessary information, and prevent information loss. Second, a new partitioning strategy is proposed for the sampled regions, which is used to enhance the connection between local information and global information. We proved that the Top-1 accuracy of the NEW-STGCN-CA model in the NTU-RGB+D 60 dataset reached 84.86%, which was 1.7% higher than the original model; the accuracy of Top-1 on the Kinetics-Skeleton dataset reached 32.40%, which was 3.17% higher than the original model. The experimental results show that NEW-STGCN-CA can effectively improve the algorithm's accuracy while also having high robustness and performance.

Keywords: deep learning; human action recognition; ST-GCN; attention mechanism; skeleton recognition

check for updates

Citation: Guo, K.; Wang, P.; Shi, P.; He, C.; Wei, C. A New Partitioned Spatial–Temporal Graph Attention Convolution Network for Human Motion Recognition. *Appl. Sci.* **2023**, *13*, 1647. https://doi.org/10.3390/app13031647

Academic Editor: Javier Hernando

Received: 29 December 2022
Revised: 10 January 2023
Accepted: 10 January 2023
Published: 28 January 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Human action recognition plays a very important role in computer vision [1]. With the development of deep learning, the traditional algorithm [2] improves the performance of the skeletal action recognition algorithm. However, it ignores the information between the nodes, and the improvement in overall performance is limited. Liu et al. [3] proposed to use a self-attention mechanism in a graph convolution network. This method increases the flexibility of the graph convolution network. However, the recognition effect will have errors when the skeleton is blocked. Zhu et al. [4] proposed an adaptive graph convolutional network. It integrates an attention mechanism into the network. However, the accuracy of this network is slightly worse than that of the existing mainstream network. Yang et al. [5] proposed the NST-GCN model. A new partition strategy is proposed, but the network cannot distinguish between similar behaviors, and the accuracy of identifying similar behaviors is poor. Song et al. [6] were able to obtain dynamic data information about the skeleton when combined with an attention mechanism so as to focus on each frame of bone data in varying degrees. However, this method only focuses on local features and ignores the relationship between the global and the local.

To sum up, ST-GCN has a lot of research results in human action recognition. Compared with traditional neural networks, it can improve recognition performance. However, ST-GCN [7] is less effective in identifying similar and local actions [8]. The model training and inference will also take up a lot of computational resources at this stage.

Our main contributions are summarized as follows.

(1) On the basis of the ST-GCN network, the CA attention mechanism [9] module is integrated into the network structure to assign appropriate weights to different joints, and the network after the introduction is optimized.
(2) The sampling region is divided in a new way. This method increases the sampling distance, and the root node is closely connected with the farther node.
(3) Finally, through a large number of experiments and analyses, it is proved that the accuracy of our model is improved compared with the existing model and that the robustness of the algorithm is improved.

The content of this paper is distributed as follows: Section 2 introduces related work. In Section 3, ST-GCN is described in detail. The method for improving ST-GCN is introduced in Section 4. The experimental results and analysis of the proposed method are presented in Section 5. Finally, in Section 6, we summarize our work.

2. Related Work

The current research shows that human action recognition based on skeletons is a topical research issue [10]. It has important applications on many occasions, such as intelligent interaction, video surveillance, medical health, and so on.

Convolutional neural networks (CNNs) [11–16] and recurrent neural networks (RNNs) [17–20] have improved the performance of action recognition algorithms based on skeleton data, which can extract joint point features from skeleton sequences. However, because skeleton data are non-Euclidean structural data [21], such algorithms ignore the crucial information related to nodes, so the overall improvement is limited. Yang et al. [22] proposed a vision-based automated method for surface condition identification of concrete structures. The recognition of concrete structures combines CNN, transfer learning, and decision-level images, and an improved Dempster–Shafer (D-S) algorithm is also designed to improve the accuracy of recognition. Shahroudy et al. [23] put forward P-LSTM, which can regard all bone points as a whole to learn human actions and add confidence gates to reduce the impact of noise information on the human recognition process. Yang et al. [24] studied the application of deep learning technology in the evaluation of the torsional bearing capacity of reinforced concrete beams, established a data-driven model based on a convolution network, and used an improved bird swarm algorithm (IBSA) to optimize the hyperparameters.

The difficulties that previous algorithms had to deal with have been solved with the emergence of graph neural networks (GNNs) [25]. It recognizes human motion by extracting the associated features between nodes. Compared with other methods, GNN is more competent for action recognition tasks based on skeleton data. At present, the mainstream method based on video and RGB (red, green, blue) images is the two-stream network [26], while the mainstream method based on skeleton data is the graph convolution network (GCN) [27]. GCN is formed by extending the concept of CNN to GNN, which combines the advantages of both to improve the accuracy of the algorithm. The spatial–temporal graph convolutional neural network (ST-GCN) is proposed based on GCN. It can extract spatial–temporal features from spatial and temporal dimensions, greatly improving the accuracy of action recognition. After considering the long-term dependency between joints, Zhang et al. [28] proposed to integrate the context information of each vertex and simplify the network structure through context-aware graph evolution (CA-GCN). Lee et al. [29] proposed a self-attention graph pooling method that uses the self-attention mechanism of graph convolution to distinguish the nodes that should be discarded and retained and can use relatively few parameters to learn end-to-end. Sun et al. [30] proposed a module integrating multilayer perceptron (MLP) and GCN, which can automatically determine key points through graph concern expression and learn in a weakly supervised way, saving time and improving the accuracy of recognition. Spadon et al. [31] proposed the recurrent graph evolution neural network (REGENN), which can simultaneously predict multiple multivariate time series. Zhang et al. [32] proposed a graph-aware transformer

(GAT), which can learn different spatial–temporal motion features from skeleton graph sequences in a data-driven manner.

3. ST-GCN

The ST-GCN can identify actions based on skeleton sequences by extending graph neural networks to the spatial–temporal graph model. It can use a graph convolution network to implicitly combine position information with temporal dynamic information. The process of a constructing a skeletal spatial–temporal graph involves taking video sequences as input. First, the input video is processed. The input video uses OpenPose [33] to estimate the human posture, so as to obtain the human skeleton sequence. After that, the recognized skeleton sequence is input to ST-GCN for training. Finally, the classification of human activities is realized. The network identification process is shown in Figure 1.

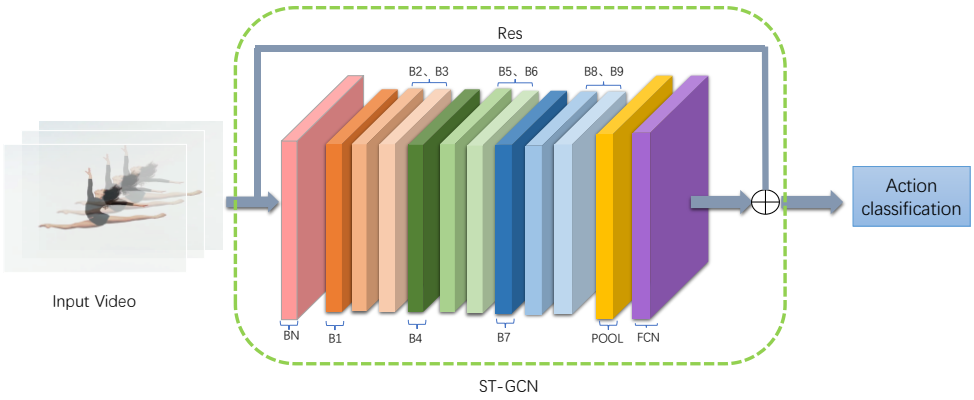


Figure 1. The human action recognition process of ST-GCN.

3.1. Constructing Skeleton Graph

The information in the skeleton spatial–temporal graph includes the number of human joints N and frames contained in the input video sequence T. A skeletal spatial–temporal graph is constructed with $G = (V, E)$ as the input to the ST-GCN; the set E represents the edge of the skeleton graph. Thus, the set of matrix eigenvalues of N joints under T frames is denoted by $V = \{v_{ti} | t = 1, \ldots, T, i = 1, \ldots, N\}$ [34], where v_{ti} is the upper i-th skeletal point of the t-th frame.

Two types of edges exist in the skeletal spatial–temporal graph [35], one of which is the spatial edge $E_S = \{v_{ti}v_{tj} | (i, j) \in H\}$, which corresponds to the natural connection of joints. The other is the temporal edge $E_F = \left\{v_{ti}v_{(t+1)i}\right\}$, which connects the same joints in successive time steps, where v_{tj} is the upper j-th skeletal point of the t-th frame. On this basis, a multilayer spatial–temporal graph convolution is constructed, which can make the information merge along the spatial and temporal dimensions. Figure 2 is the constructed spatial–temporal graph.

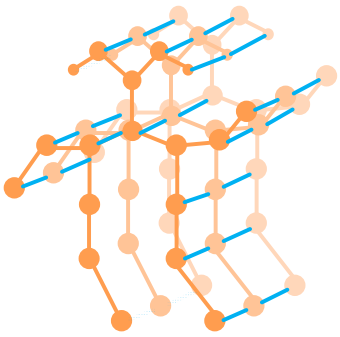


Figure 2. The spatial–temporal graph. Orange dots denote body joints. The blue line represents the inner edge between the naturally connected body joints.

3.2. Constructing Spatial–Temporal Graph Convolution Kernels

The human skeleton sequence is connected by video frames, and each frame is a 2D image containing the skeleton graph of the human body [36]. The skeleton edges $E_S(\tau) = \{v_{ti}v_{tj}, (i,j) \in H\}$ are composed of a single frame τ and V_t with N joint nodes. After the image in the convolutional neural network is convolved, its output is still a 2D grid. The output value for a single channel at the spatial location x can be written as Equation (1).

$$f_{out}(x) = \sum_{h=1}^{K} \sum_{\omega=1}^{K} f_{in}(p(\mathrm{x}, h, \omega)) \cdot \omega(h, \omega) \tag{1}$$

where f_{in} is an input feature map with the number of channels c, the sampling function $p(h, \omega)$ is the surrounding adjacent pixels centered on the x pixel and the size of the sampling region is the same as the filter size; ω is the weight function matrix.

$f_{out}(x)$ provides a weight vector in the c-dimensional space, which is used to calculate the inner product with the input characteristic matrix. The graph convolution formula can be constructed by redefining the sampling function p and the weighting function ω. The set of pixels in the adjacent region is represented by $B(v_{ti}) = \{v_{tj} | d(v_{tj}, v_{ti}) \leq D\}$, where $d(v_{tj}, v_{ti})$ denotes the shortest distance from v_{tj} to v_{ti}. The sampling function (D = 2) can be expressed by Equation (2). In 2D convolution, the adjacent-region pixels are regularly arranged around the central pixel, so they can be convolved with a regular convolution kernel based on spatial order. The adjacent-region pixels obtained by the sampling function are divided into different subsets, each of which corresponds to a numerical label. Thus, $l_{ti} : B(v_{ti}) \rightarrow \{0, \ldots, K-1\}$ maps the nodes in the adjacent set to the corresponding subset labels. So, the new weight function is represented by Equation (3). The Equation (4) is the convolution formula of the spatial graph.

$$p(v_{ti}, v_{tj}) = v_{tj} \tag{2}$$

$$w(v_{ti}, v_{tj}) = w'(l_{ti}(v_{tj})) \tag{3}$$

$$f_{out}(v_{ti}) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(v_{tj}) \cdot w'(l_{ti}(v_{tj})) \tag{4}$$

where $Z_{ti}(v_{tj}) = |\{v_{tj} | l_{ti}(v_{tk}) = l_{ti}(v_{tj})\}|$ is a normalization term.

The temporal dimension is described by connecting two identical skeletal points in two adjacent frames when constructing the skeleton spatial–temporal graph. After selecting a skeleton frame sequence with a time range of Γ, spatial graph convolution is applied within this range. The spatial graph convolution formula of Equation (4) can be extended to the time dimension, so that the spatial–temporal graph convolution formula is Equation (5). In the temporal dimension, the functions of the sampling and weight are used to define the mapping result l_{ST} of bone nodes in adjacent regions. The mapping result is Equation (6). This can realize convenient spatial–temporal graph convolution operations on a skeleton spatial–temporal graph.

$$B(v_{ti}) = \{v_{qj} | d(v_{tj}, v_{ti}) \leq K, |q - t| \leq \Gamma/2\} \tag{5}$$

$$l_{ST}(v_{qj}) = l_{ti}(v_{tj}) + (q - t + \Gamma/2) \times K \tag{6}$$

where Γ controls the convolution kernel in the temporal domain, and $l_{ti}(v_{tj})$ is the label map at v_{tj}.

3.3. The Way the Sampling Area Is Divided

The domain of each node is divided into regions for spatial graph convolution. ST-GCN has three partitioning strategies, as shown in Figure 3. Body joints are indicated by orange dots, and in Figure 3, red dashed circles are the adjacent area $B(v_{ti})$ with the sampling function p using distance D = 1.

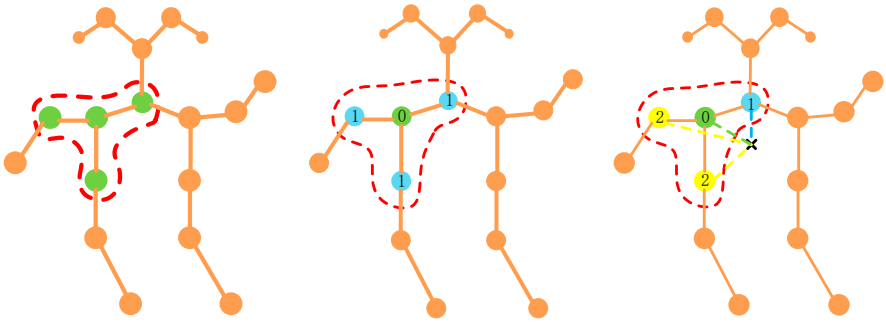


Figure 3. The way the sampling area is divided. (**a**) Unilabeling partitioning strategy. All nodes in the neighborhood have the same label (green); (**b**) distance partitioning strategy. These two subsets are the root node itself with a distance of 0 (green) and other adjacent points with a distance of 1 (blue).; (**c**) spatial configuration partitioning. Center of gravity of the skeleton (black cross), centripetal node (blue), and centrifugal node (yellow).

The unilabeling partitioning strategy is depicted in Figure 3a. The adjacent region $B(v_{ti})$ of root node v_{ti} is not divided.

Figure 3b is the distance partitioning strategy. The adjacent region $B(v_{ti})$ of the root node v_{ti} is divided into two regions: the root node 0 (green dots) and the remaining neighbor nodes (blue dots). Each region is assigned a different weight.

The spatial configuration partitioning strategy is shown in Figure 3c. This strategy can be broadly classified into centripetal and centrifugal motions according to the motion of the body parts. Take the bone center of gravity as the reference point, which is indicated by a black cross. The $B(v_{ti})$ is divided into three regions: (1) the root node itself (green dot); (2) the centripetal set (blue dot)—the neighboring nodes that are closer to the gravity center of the skeleton than the root node; (3) the centrifugal set (yellow dots)—the centrifugal group. Each region is assigned a different weight.

4. Optimization of ST-GCN Neural Networks

In this study, ST-GCN is optimized to improve the detection results since its attention mechanism lacks flexibility and only focuses on local information: (1) The attention mechanism is integrated into the ST-GCN network structure and appropriate weights are allocated. The model is named STGCN-CA; (2) A new partitioning strategy is proposed for the sampling region, which closely connects the root node with the farther node. This method can enhance the transmission of information between the whole and the local, and it is named NEW-STGCN; (3) The model after merging the two optimization schemes is named NEW-STGCN-CA.

4.1. CA Attention Mechanism Module

The remarkable advantage of attention mechanisms is that they pay attention to relevant information while ignoring irrelevant information, establishing direct dependency between input and output without the need for circulation. The degree of parallelism is enhanced, and the running speed is greatly improved. It overcomes some limitations of traditional neural networks, such as the performance degradation of the system with the increase in the input length, the low computational efficiency of the system due to the unreasonable input order, and the lack of feature extraction and enhancement of the system. However, attention mechanisms can well model serial data with variable length, further enhancing their ability to capture remote dependency information, reduce the depth of hierarchy, and effectively improve accuracy.

At present, the most popular attention mechanism is the squeeze-and-excitation (SE) [37] attention mechanism. With the help of the global pooling layer, it calculates the channel information attention and realizes the gain function at a low cost. However, SE

only considers the information between channels and ignores the importance of location information. Secondly, the convolutional block attention module (CBAM) [38] obtains location information by reducing the channel dimension of the input tensor and using convolution to calculate spatial attention.

The CA attention mechanism incorporates location information into channel data. Because of these advantages, it helps to improve the positioning and recognition abilities of the model. Then, the CA attention mechanism is adaptable and lightweight, allowing it to be easily integrated into a mobile network (for example, MobileNetV2 [39], MobileNeXt [40], and EfficientNet [41]). It can enhance features by emphasizing information representation while avoiding a large amount of computation.

The information is embedded in a global method. However, the global method is usually used for the global coding of channel coding spatial information, and it is difficult to save the location information. To address this problem, the global pooling is decomposed according to Equation (7) and transformed into a pair of one-dimensional feature encoding operations. Next, each channel is coded by the pooling kernels (H, 1) or (1, W). Therefore, the output of the c-th channel with height h is represented as Equation (8) and the output of the c-th channel with width w is Equation (9).

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j) \tag{7}$$

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, j) \tag{8}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \tag{9}$$

where z_c is the output associated with the c-th channel.

The features of height and width are gathered, and a pair of feature maps with direction perception are obtained. Therefore, the attention mechanism can capture the channel space information in one direction and obtain the position information in the other direction. Using the 1×1 convolutional transform function for F_1, its transform is shown in Equation (10). Then, f is decomposed into two separate tensors: $f^h \in R^{\frac{C}{r} \times H}$ and $f^w \in R^{\frac{C}{r} \times w}$. The other 1×1 convolutional transforms, F_h and F_w, transform tensors with the same number of channels into the input. The attention is weighted as Equations (11) and (12), respectively.

$$\mathrm{f} = \delta\left(F_1\left(\left[\mathrm{z}^h, \mathrm{z}^w\right]\right)\right) \tag{10}$$

$$\mathrm{g}^h = \sigma\left(F_h\left(\mathrm{f}^h\right)\right) \tag{11}$$

$$\mathrm{g}^w = \sigma(F_w(\mathrm{f}^w)) \tag{12}$$

where $[\cdot, \cdot]$ is a concatenation operation along the spatial dimension, δ is a nonlinear activation function, and f is an i feature mapping. σ is the sigmoid activation function. The outputs g^h and g^w are expanded as attention weights, respectively. Finally, the output Y of the CA module can be written as in Equation (13).

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \tag{13}$$

In this study, the attention mechanism will focus on the extraction of human motion features. We propose the STGCN-CA attention model, which incorporates the CA attention into the ST-GCN. After optimizing the network, the STGCN-CA network structure is shown in Figure 4.

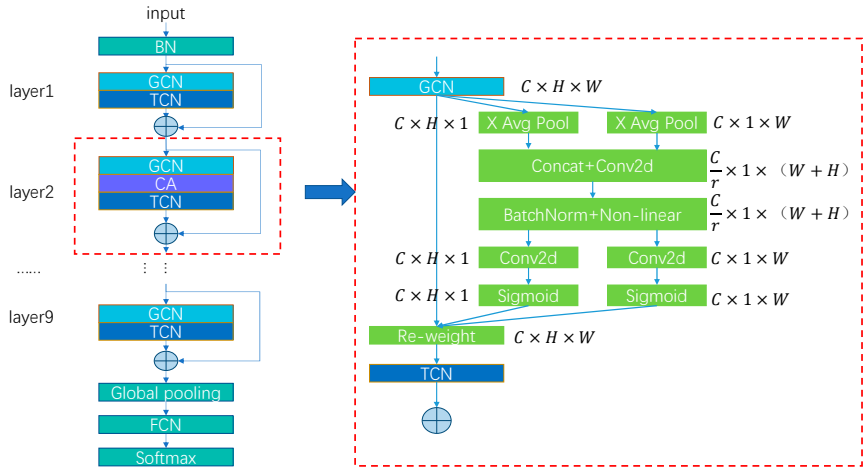


Figure 4. Network structure diagram of STGCN-CA.

Firstly, the STGCN-CA network performs global averaging pooling of the feature map output from the GCN layer along both the width and height directions. It is decomposed of two parallel one-dimensional feature codes to obtain feature maps in width and height directions. These feature maps are spliced together and sent to the shared convolution module for dimension reduction. Secondly, the feature map after batch normalization is sent into the Sigmoid activation function, and the feature maps with the same number of original channels are obtained by convolution according to the original height and width. Then, we use the Sigmoid activation function to calculate the attention weight in the height and width directions, respectively. Finally, the original feature maps are multiplied and weighted to produce the attention-weighted feature maps.

4.2. Constructing a New Partition Strategy

A good partition strategy can significantly improve model feature extraction capability. In order to make up for the lack of global information, a new partition strategy is used in this paper. We set the sampling function p to use the adjacent region with distance D = 2. It is divided into four subsets: (1) the root node itself (green dot); (2) the centripetal group (blue dot)—skeletal points that are closer to the center of gravity of the skeleton than the root node and that are adjacent to the root node; (3) the centrifugal group (yellow dots)—skeletal points further away from the center of gravity of the skeleton than the root node and adjacent to the root node; (4) the telecentric group (pink dots)—nodes at D = 2 from the root node. The partitioning strategy NEW-STGCN is shown in Figure 5. The purple dotted line part is the newly proposed extension region.

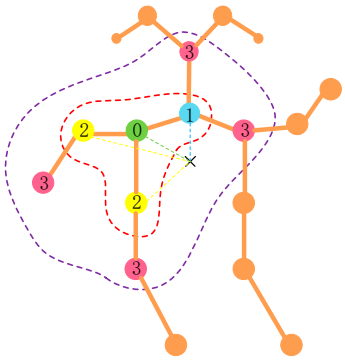


Figure 5. New partitioning strategy. The root node is represented by a green dot, the centripetal point is represented by a blue dot, the yellow dot represents a centrifugal dot, and the pink dot represents a point 2 from the root node.

This partitioning strategy expands the sampling region to consider the movement form of local joints as well as the mutual interconnection between joints. It can strengthen the link between the root node and the farther node by connecting information from all parts of the body, allowing the model to better perceive local body information and improve recognition accuracy.

5. Experimental Results

5.1. Data Description

The NTU-RGB+D 60 dataset is a large dataset for human activity analysis. It contains 60 action categories with a total of 56,880 RGB+D video samples using Microsoft Kinect V2. Some sample data from the NTU RGB+D 60 dataset are shown in Figure 6. Every action video should not exceed two human skeleton sequences, and there are 25 key joints in each skeleton sequence.



Figure 6. NTU RGB+D 60 dataset partial sample data.

The NTU-RGB+D 60 dataset provides two subsets: (1) Cross-Subject (CS) is a dataset based on different behavioral subjects. It divided the daily action videos of 40 volunteers into two training groups and test groups. There were 40,320 videos for training samples and 16,560 videos for testing samples in the other group. (2) Cross-View (CV) is a dataset based on different perspectives. It uses three cameras to take three different views from the same location at the same time.

The Kinetics-Skeleton dataset is based on the Kinetics (Kinetics human action dataset) large-scale motion recognition dataset. Kinetics contains about 300,000 video clips retrieved from YouTube, covering up to 400 human actions, and each clip in the video lasts about 10 s.

The evaluation indicators are $Top-1$ *accuracy* and $Top-5$ *accuracy*. $Top-1$ *accuracy* means the possibility that the category ranking first is consistent with the actual results; $Top-5$ *accuracy* is the accuracy of the top five categories containing the actual result. These equations are as follows:

$$Top-1\ accuracy = \frac{N}{M} \tag{14}$$

$$Top-5\ accuracy = \frac{Y}{M} \tag{15}$$

where N is the sample size of all test images with the best probability of being correctly labeled; M is the total sample size; and Y is the sample size of all test images with the top five probabilities of being correctly labeled.

5.2. Training Details

All experiments were run on a computer with one NVIDIA GeForce GTX 1080Ti and an Intel(R) Core (TM) i7-7700 running at 3.60 GHz. The PyTorch deep learning framework was built with Ubuntu 16.04 LTS architecture. The program code was written using Python language and called CUDA, CUDNN, and other required libraries to achieve the training and testing of human motion recognition models. The improved ST-GCN was trained by stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay of 0.0001. There were 50 epochs in our training process, with 13 steps per epoch, and the complete training took about 600 min. The batch size was set to 16 and the learning rate was set to 0.1. When epochs achieved 10 and 50, the learning rate was reduced by a factor of 0.1.

5.3. Experimental Comparison and Analysis

To verify the effectiveness of this study, the proposed optimization algorithm was subjected to ablation experiments. Since human action recognition is a multiclassification task, the cross-entropy loss function was used to train the model. According to the training data, a line chart of the change in loss function value with the iteration number was drawn, as shown in Figure 7.

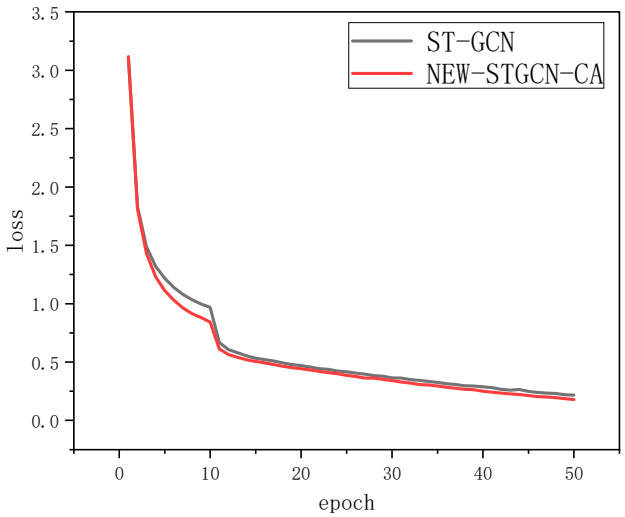


Figure 7. Change curve of loss values.

On the CS benchmark, the comparison of loss changes obtained by NEW-STGCN-CA and ST-GCN is shown in Figure 7, where both models converge. The NEW-STGCN-CA model converged faster than the ST-GCN model from the start of training until the epoch reaches 10, and the loss value was lower in the same epoch. Then the two curves gradually decreased, and the overall loss value of NEW-STGCN-CA was lower than ST-GCN.

First of all, we integrated three different attention mechanisms (SE, CBAM, and CA) into the same position of the ST-GCN and compared them with the experimental results of the original model. According to the experimental results, ST-GCN-CA with the CA attention mechanism has the best experimental results. The experimental results in Table 1 show that the introduction of an attention mechanism can ensure that more discriminant features are assigned higher weights in the calculation process.

Table 1. Experimental results on the NTU-RGB+D 60 dataset.

Algorithms	CS	CV
ST-GCN	81.69%	88.30%
ST-GCN+SE	84.15%	89.30%
ST-GCN+CBAM	83.90%	89.30%
STGCN-CA	84.45%	90.10%
NEW-STGCN	83.49%	91.60%
NEW-STGCN-CA	84.86%	92.46%
Two-stream	83.20%	89.30%
Clip+CNN+MTLN	79.60%	84.80%
ARRN-LSTM	81.80%	88.00%
BPLHM	84.50%	91.10%
CA-GCN	83.50%	91.40%

Adaptive focusing is carried out with different human behaviors to distinguish key nodes and improve the accuracy of recognition. Compared with ST-GCN, NEW-STGCN increases the number of partitions in the sampling region. With the increase in convolution kernels, the network can learn more abundant human joint connections. Therefore, the accuracy of identification can be improved.

It can be seen from Table 1 that the accuracy of the NEW-STGCN-CA scheme is improved compared with ST-GCN. With the CS benchmark, the accuracy of the NEW-STGCN-CA algorithm is 3.16% higher than that of the ST-GCN algorithm. On the CV benchmark, NEW-STGCN-CA's accuracy is better by 4.16%. The CV benchmark is the video captured by the 3-angle cameras, which can better extract features and have a high recognition rate. On the CS benchmark, the results show that different recognition objects have great influence on recognition.

On the NTU RGB+D 60 dataset, we compared the recognition effect of the NEW-STGCN-CA with several typical methods to validate the feasibility of this research method. First of all, using the classic two-stream model, our accuracy on the CS benchmark is 1.66% higher than two-stream's and 3.16% higher on the CV benchmark. The Clip+CNN+MTLN model is improved based on CNN. Compared with the NEW-STGCN-CA, the accuracy is improved by 5.26% on the CS benchmark and 7.66% on the CV benchmark. On the NTU-RGB+D dataset, the NEW-STGCN-CA model is 3.06% higher in the CS benchmark than the improved ARRN-LSTM model based on RNN and 4.46% higher in the CV benchmark. The accuracy of NEW-STGCN-CA on the CS and CV benchmarks is 0.36% and 1.36% higher than that of BPLHM, respectively. The performance of NEW-STGCN-CA on NTU-RGB+D is 1.36% better than that of CA-GCN on the CS benchmark and 1.06% better than that of the CV benchmark.

Similarly, on the Kinetics-Skeleton dataset, we compared the recognition effect of NEW-STGCN-CA with four typical methods, namely ST-GCN, feature Enc based on feature coding, and Deep LSTM and Temporary ConvNet based on deep learning. The experimental results are shown in Table 2. NEW-STGCN-CA is superior to these four typical methods in two evaluation indicators. The accuracy of Top-1 and Top-5 in NEW-STGCN-CA reached 32.40% and 54.80%, respectively. Compared with ST-GCN, the Top-1 accuracy and Top-5 accuracy of NEW-STGCN-CA are 1.7% and 2% higher, respectively. Compared with Feature Enc, Top-1 accuracy and Top-5 accuracy have improved by 17.5% and 29%, respectively. The accuracy of NEW-STGCN-CA is 16% and 19.50% higher than that of the Deep LSTM Top-1 accuracy and Top-5 accuracy, respectively, and 12.1% higher than Top-1 accuracy and 14.80% higher than Top-5 accuracy for Temporary ConvNet. The reason for the different recognition rates on the Kinetics-Skeleton and NTU RGB+D 60 datasets is that the recognition rate on Kinetics-Skeleton is low due to the harsh data collection environment.

Table 2. Comparison of recognition accuracy of Kinetics-Skeleton dataset.

Algorithms	Top-1	Top-5
NEW-STGCN-CA	32.40%	54.80%
Deep LSTM	16.40%	35.30%
Temporal Conv	20.30%	40.00%
Feature Enc	14.9%	25.80%
ST-GCN	30.7%	52.80%

The NEW-STGCN-CA can accurately identify the actions in the target video in Figure 8. However, the existing two-stream has misjudgment and cross-recognition effects. For example, tai chi is recognized as baby crawling in Figure 8a, and the stretching leg action is recognized as snowboarding in Figure 8b.

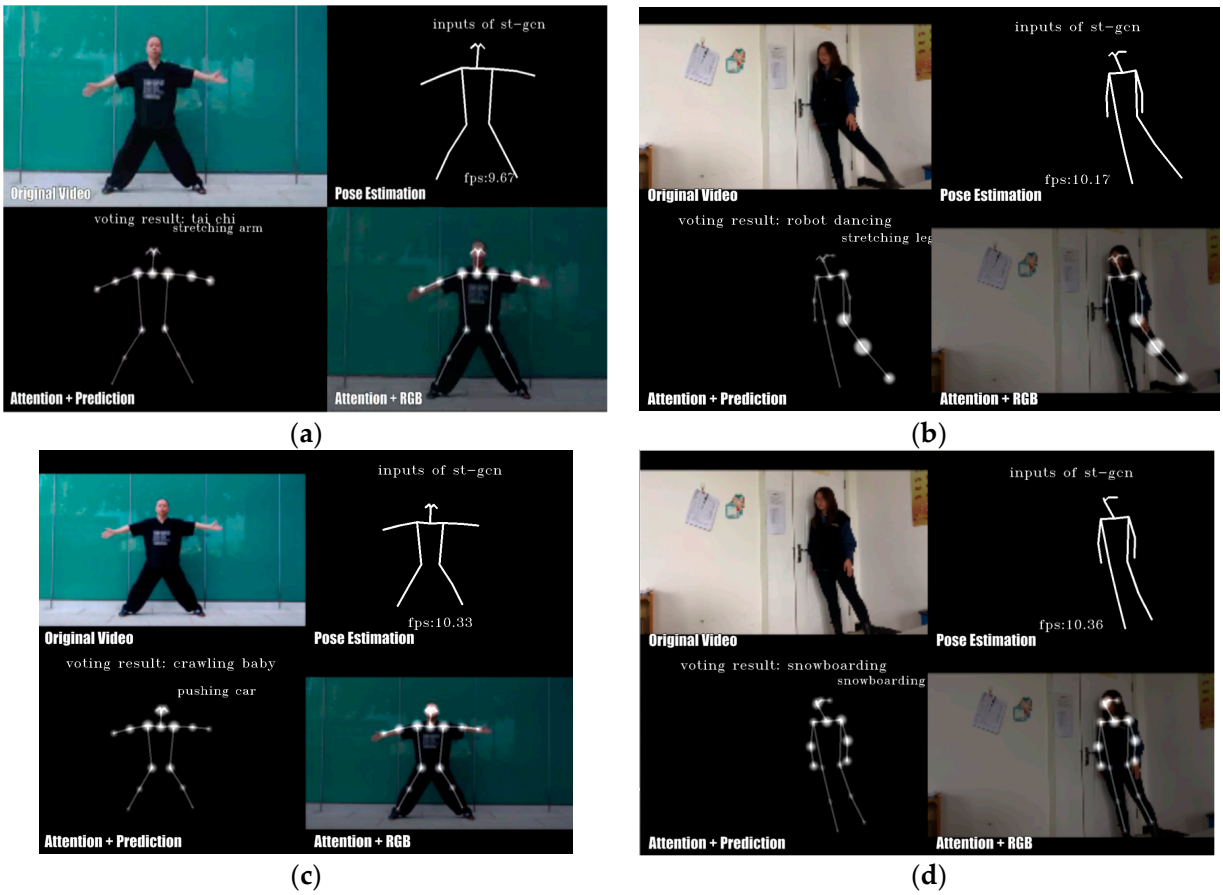


Figure 8. Comparison of NEW-STGCN-CA and two-stream identification. Figures (**a**,**c**) show NEW-STGCN-CA and Two-Stream recognizing the stretching arm action in Tai Chi, and Figures (**b**,**d**) show NEW-STGCN-CA and Two-Stream recognizing the stretching leg action, respectively.

In practical applications, images may be contaminated by different noise sources. Therefore, we enhanced the data of the sample image, added different noise processing as shown in Figure 9, and then used NEW-STGCN-CA to identify the actions in the image. The recognition effect is shown in Figure 10. The experimental results show that the recognition effect of NEW-STGCN-CA in raising arm movements in tai chi movements is accurate under the interference of noise.

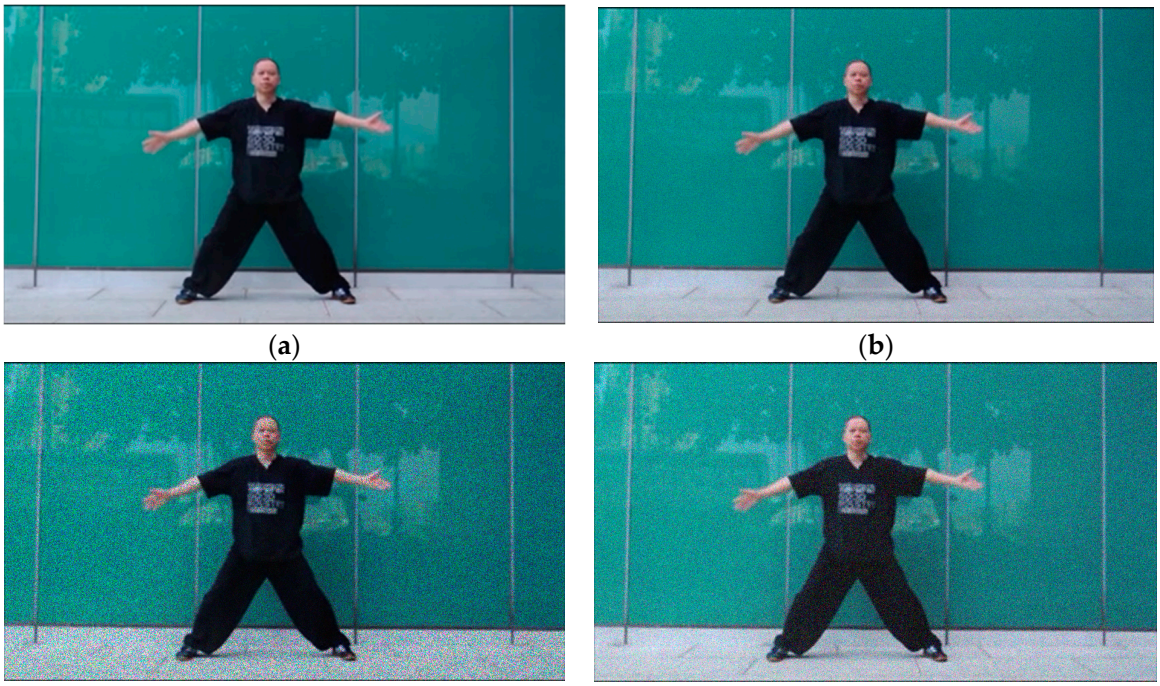


Figure 9. Data enhancement. Noise was added to the sample image. Figure (**a**) is the original image, Figure (**b**) adds Poisson noise, Figure (**c**) adds Speckle noise, and Figure (**d**) adds Gaussian noise.

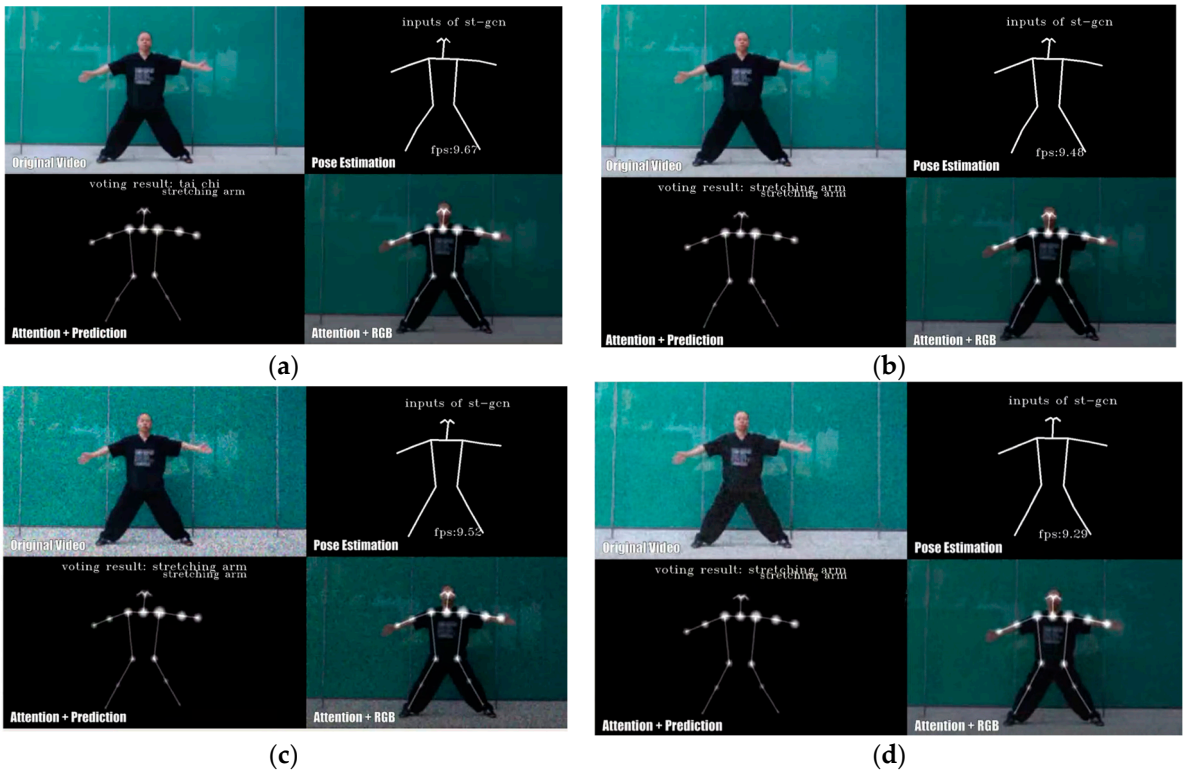


Figure 10. NEW-STGCN-CA recognition effect of noise sample image. Figure (**a**) is the original image, Figure (**b**) adds Poisson noise, Figure (**c**) adds Speckle noise, and Figure (**d**) adds Gaussian noise.

6. Conclusions

In this paper, we propose a NEW-STGCN-CA model for human action recognition. First of all, the CA attention mechanism is integrated between the spatial graph layer and

the temporal graph layer. It solves the problem that key nodes cannot be distinguished for different behaviors and improves the accuracy of model identification. Secondly, a new partitioning strategy is proposed for the sampled region. It strengthens the connection between the local information and the global information, and improves the robustness of the algorithm. On the NTU RGB+D 60 dataset, we first carried out the ablation experiment of the attention mechanism, and the results showed that the recognition accuracy of STGCN-CA was better than ST-GCN and the other two attention mechanisms. Secondly, the new partition strategy and CA attention mechanism are integrated into ST-GCN at the same time. The recognition accuracy of NEW-STGCN-CA is better than that of ST-GCN. At the same time, the recognition accuracy of NEW-STGCN-CA has improved to a certain extent compared with other mainstream networks on NTU RGB+D 60 and Kinetics-Skeleton datasets. The next step is to transplant this algorithm into raspberry pie, use the image read by the camera to recognize the human body in the monitoring area, further analyze the effect in practical application, and continuously optimize it.

Author Contributions: Contributed to the experimental equipment, K.G.; performed the experiments, wrote the original draft, and analyzed the results, P.W.; writing—review and editing, P.S.; writing—review and editing, C.H.; writing—review and editing, C.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data or code presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hu, W.; Tan, T.; Wang, L.; Maybank, S. A Survey on Visual Surveillance of Object Motion and Behaviors. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2004**, *34*, 334–352. [CrossRef]
2. Ravanbakhsh, M.; Mousavi, H.; Rastegari, M.; Murino, V.; Davis, L.S. Action Recognition with Image Based CNN Features. *arXiv* **2015**, arXiv:1512.03980.
3. Liu, C.; Fu, R.; Li, Y.; Gao, Y.; Shi, L.; Li, W. A Self-Attention Augmented Graph Convolutional Clustering Networks for Skeleton-Based Video Anomaly Behavior Detection. *Appl. Sci.* **2021**, *12*, 4. [CrossRef]
4. Zhu, Q.; Deng, H.; Wang, K. Skeleton Action Recognition Based on Temporal Gated Unit and Adaptive Graph Convolution. *Electronics* **2022**, *11*, 2973. [CrossRef]
5. Yang, S.; Li, Q.; He, D.; Wang, J.; Li, D. Global Correlation Enhanced Hand Action Recognition Based on NST-GCN. *Electronics* **2022**, *11*, 2518. [CrossRef]
6. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. Spatio-Temporal Attention-Based LSTM Networks for 3D Action Recognition and Detection. *IEEE Trans. Image Process.* **2018**, *27*, 3459–3471. [CrossRef]
7. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018. [CrossRef]
8. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
9. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021. [CrossRef]
10. Perrot, J.Y.; Boucheix, C.; Mirshahi, M.; Kazatchkine, M.; Bariety, J. Monoclonal antibodies against surface antigens of lymphoblasts and blood cells or bone marrow recognize constituents of the human nephron. *Nephrologie* **1984**, *5*, 53–57.
11. Kim, T.S.; Reiter, A. Interpretable 3D Human Action Analysis with Temporal Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017.
12. Bo, L.; Dai, Y.; Cheng, X.; Chen, H.; He, M. Skeleton based action recognitionSkeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China, 10–14 July 2017.

13. Li, C.; Hou, Y.; Wang, P.; Li, W. Multiview-Based 3-D Action Recognition Using Deep Networks. *IEEE Trans. Hum.-Mach. Syst.* **2019**, *49*, 95–104. [CrossRef]
14. Yang, F.; Wu, Y.; Sakti, S.; Nakamura, S. Make Skeleton-based Action Recognition Model Smaller, Faster and Better. In Proceedings of the ACM Multimedia Asia 2019, Beijing, China, 16–18 December 2019. [CrossRef]
15. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. A New Representation of Skeleton Sequences for 3D Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
16. Cao, C.; Lan, C.; Zhang, Y.; Zeng, W.; Lu, H.; Zhang, Y. Skeleton-Based Action Recognition With Gated Convolutional Neural Networks. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 3247–3257. [CrossRef]
17. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016. [CrossRef]
18. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1110–1118.
19. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017. [CrossRef]
20. Li, S.; Li, W.; Cook, C.; Zhu, C.; Gao, Y. Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–22 June 2018. [CrossRef]
21. Laub, J.; Roth, V.; Buhmann, J.M.; Müller, K.-R. On the information and representation of non-Euclidean pairwise data. *Pattern Recognit.* **2006**, *39*, 1815–1826. [CrossRef]
22. Yu, Y.; Samali, B.; Rashidi, M.; Mohammadi, M.; Nguyen, T.N.; Zhang, G. Vision-based concrete crack detection using a hybrid framework considering noise effect. *J. Build. Eng.* **2022**, *61*, 105246. [CrossRef]
23. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
24. Yu, Y.; Liang, S.; Samali, B.; Nguyen, T.N.; Zhai, C.; Li, J.; Xie, X. Torsional capacity evaluation of RC beams using an improved bird swarm algorithm optimised 2D convolutional neural network. *Eng. Struct.* **2022**, *273*, 115066. [CrossRef]
25. Velikovi, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. *arXiv* **2017**, arXiv:1710.10903.
26. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. *Adv. Neural Inf. Process. Syst.* **2014**, *1*, 1–9.
27. Xu, M.; Zhao, C.; Rojas, D.S.; Thabet, A.; Ghanem, B. G-TAD: Sub-Graph Localization for Temporal Action Detection, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, Seattle, WA, USA, 13–19 June 2020. [CrossRef]
28. Zhang, X.; Xu, C.; Tao, D. Context Aware Graph Convolution for Skeleton-Based Action Recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
29. Lee, J.; Lee, I.; Kang, J. Self-Attention Graph Pooling. In Proceedings of the International Conference on Machine Learning 2019, Long Beach, CA, USA, 10–15 June 2019.
30. Sun, L.; Zhang, Z.; Zhong, R.; Chen, D.; Zhang, L.; Zhu, L.; Wang, Q.; Wang, G.; Zou, J.; Wang, Y. A Weakly Supervised Graph Deep Learning Framework for Point Cloud Registration. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5702012. [CrossRef]
31. Spadon, G.; Hong, S.; Brandoli, B.; Matwin, S.; Rodrigues, J.F., Jr.; Sun, J. Pay Attention to Evolution: Time Series Forecasting with Deep Graph-Evolution Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 5368–5384. [CrossRef] [PubMed]
32. Zhang, J.; Xie, W.; Wang, C.; Tu, R.; Tu, Z. Graph-aware transformer for skeleton-based action recognition. *Vis. Comput.* **2022**, 1–12. [CrossRef]
33. Chen, C.H.; Ramanan, D. 3D Human Pose Estimation = 2D Pose Estimation + Matching. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
34. Zhe, C.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
35. Peng, Y.; Zhao, Y.; Zhang, J. Two-Stream Collaborative Learning with Spatial-Temporal Attention for Video Classification. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 773–786. [CrossRef]
36. Das, P.P. Human skeleton tracking from depth data using geodesic distances and optical flow. *Comput. Rev.* **2013**, *54*, 702.
37. Jie, H.; Li, S.; Gang, S.; Albanie, S. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; p. 99.
38. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. *CBAM: Convolutional Block Attention Module*; Springer: Cham, Switzerland, 2018. [CrossRef]
39. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018. [CrossRef]

40. He, X.; Cheng, R.; Zheng, Z.; Wang, Z. Small Object Detection in Traffic Scenes Based on YOLO-MXANet. *Sensors* **2021**, *21*, 7422. [CrossRef] [PubMed]
41. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the International Conference on Machine Learning 2019, Long Beach, CA, USA, 10–15 June 2019.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.